A Program for Better Human Computer Collaboration to Counter Terrorism

Paul B Kantor[1] (Rutgers), Catherine L. Smith (Kent State University), Ying Sun (University at Buffalo)

Theoretical social sciences promise understanding and generalizability. They may illuminate causes of radicalization, and relations to socio-economic and geo-political trends. We address some applied aspects of the social sciences, which may strengthen the processes used to discover and defeat security threats. Specifically, we seek information systems that work optimally with their specific users, enabling several systems to work together, supporting multiple humans. Eventually, these systems will "learn from" their users, to improve the performance of both. Mathematically today's information systems are not human⊕computer[2], but human⊗computer[3]. Each component strengthens the other, and the whole is more than the sum of its parts.

Several (sub)disciplines are relevant, including "user-oriented system evaluation," "user modeling," and "usability engineering." As their names suggest, the first of these concentrates on understanding how to assess (and then improve) systems for human use; the second concentrates on understanding the user now at hand, in the context of broader theories about human computer interaction, learning styles, and cognitive individualities; the third concentrates on changing or building systems to be more usable, ideally using information gleaned from the first and second.

This note concentrates on the first of these three disciplines, user-oriented system evaluation. In the contexts of national security, information systems help to monitor and filter huge streams of information, some government-derived (from sensor and monitoring activities) and others from the exponentially growing mass of open-source and social media information. In the today's world almost anyone with a mobile phone is a potential sensor, and the information flowing among those phones and from them onto open platforms almost surely contains early hints of most adversarial actions, before they occur. This is clearly an example of a human⊗computer problem, as neither people nor computers alone can cope with the flood.

New ideas about systems are proposed almost daily. It is essential to have an orderly and principled way to assess and evaluate them. Without academic advances in system evaluation, we don't know how to assess and select evaluation methods that provide useful information for building the next generation of systems.

The National Security Agency has recognized these problems for many years, and was a key early supporter of the NIST/TREC activities. Similar programs have been developed in Europe and Asia. TREC did seek sound experimental design (Over, 2001), but those

---

[1] Corresponding author: paul.kantor@rutgers.edu

[2] "human *plus* computer."

[3] "human *times* computer." Multiplication shows that with either factor missing, the combination is almost useless. A computer alone cannot do 50 percent of the job. Nor can the analyst without her system do 50 percent of the job.

initiatives, particularly due to Pirolli (Hearst et al., 1995) at PARC, eventually succumbed to tyranny by the evaluated.

What does it mean for one system to "work better" than another? What factors might influence that performance and should be assessed during measurement? The original experiments were done in the U.K., by (Cleverdon, 1967). Key measures that remain relevant and popular to this day were systematized by (Van Rijsbergen, 1974). These include *precision (p), Recall (R), F-measures and* F_β *,* which is a kind of harmonic mean. There are two underlying probabilities, *detection (d)* of useful materials, and *false alarm (f).* But a relevant item buried beneath 20 irrelevant ones provides little value, leading to measures such as *discounted cumulated gain*. Fundamentally detection and false alarm both depend on some threshold parameter, and performance is a functional relation, *d(f)* parametrized by some threshold setting. (Swets et al., 1961; Kantor and Boros, 2010).
The performance of a system is a *relation* between the benefit and the cost or effort. A system that always has more detection at the same cost as another system, is clearly better.

What are the "variables that matter?" There are many characteristics of users as well as information objects have been named in studies that intend to understand effectiveness in real-life information seeking contexts (Saracevic & Kantor, 1988a, Saracevic & Kantor 1988b; Sun, 2005; Al-Maskari & Sanderson, 2011) Broadly: the problem; the context; the user's (knowledge; cognitive abilities; current "cognemotional" state[4]); the corpus; and, of course, the system. Thus:

RelativePerformance(A:B)=f(user(state; history); problem; context; corpus; ***A, B****)*

The comparison between two systems: ***A*** and ***B*** will depend on the user, the problem, the context, and the corpus or stream being monitored. Performance is an abstraction. If operationalized using expert judges, the corresponding *RelativeScore* depends also on the judge:

RelativeScore(A:B)=f(user(state; history); problem; context; corpus, ***judge****;* ***A, B****)*

Comparison of systems should ask: how useful is the resulting report or action? Preferably, assessments are made by the same experts whose products are being assessed. Ultimately this process could be part of normal work flows. We propose laboratories, (Kantor, 1988) at which system assessment will be done by real intelligence analysts, working on real problems of current interest.

Some variant of the General Linear Model is a promising basis for statistical analysis. The emphasis is not on the variance explained by each factor – the users will always explain the largest part. It is, rather, the partial contribution of each variable, to the overall score. In particular, once we have settled on a performance score statistics can reveal how much of the difference in the quality of analytic products, is attributable to the difference between two (or more) computer systems. Cross-evaluation (Sun and Kantor, 2006) is one

[4] We suggest this awkward term for the established fact that a person's reasoning ability may be constrained or enhanced by his or her immediate emotional status.

promising approach to the design and analysis of such studies. Cross-evaluation will also assess the individual strengths of the users, and their biases as judges.

Intelligence analysis requires the efforts of many people. They have individual styles and skills; yet must work together. In the future systems must be trained on and tuned to those people and groups who use them. The information systems (IS) will not be like cars –a few basic designs support all users. The IS for national security will be somewhat like personalized medicine and, like personalized medicine, personalized system adaptation may exploit characteristics of which the user is not consciously aware.

Realizing the full potential of these concepts from the Social Sciences will require some "instrumentation" of the IS to be evaluated. Information systems today have many parts such as "filters" and "rankers." For these systems to benefit from assessment there will be feedback from evaluations of the end product. IS must be able to trace backwards, to identify where they erred, by maintaining information about the provenance of their assertions.

If we can also know enough about the user in the "results ← user⊗problem⊗system" equation these laboratories will be able to do off-line "what-if" experiments, to improve the system. At an almost ontological level, we also need research that will operationalize conceptual classifications of people, and align them with operationalizable characteristics of systems. Only then can we match user and system for optimal security.

Specific Proposals

The social sciences have residual anti-national prejudice that emerged during the Vietnam War and metastasized into the academy; with inbred selection processes, it is slow to metabolize away. Perhaps because of this, there has been some tendency to condition support for governmental efforts on one's personal view about the party in power, or the person in the White House. In contrast to the surge of scientific input that followed World War II, and continues in the natural sciences, this polarization has limited the contributions of contemporary social scientists to important issues affecting national security. For example, the Intelligence Community (IC) Post-Doctoral program has evolved so that students have an appointment at Oak Ridge National Laboratory, and it cannot be said that they are employees of their university, while working for the IC.

How will the proposed academic research connect with the specific national problems? It might begin with pilot studies supported by the NSF, NIH; IMLS; or NIJ. Academics could propose, refine and validate needed instruments. Example "instrument problems" are: (a) what qualities of individual users (Saracevic and Kantor, 1988b) can predict how well those users will perform, regardless of the system and possibly other significant variables? (Smith and Kantor, 2008); (b) what qualities of individual users predict whether having them work "together" [in parallel, asynchronously, or elsewise] produces valuable increases in performance; and (c) what qualities of managers, and management protocols tend to enhance (synergize) the performance of groups of individual users. For example, as is known for work of TSA screeners, might cognitive workers, such as intelligence

analysts, work better with a mix of tasks, enabling them to "refresh their minds" while keeping the overall goal in focus[5].

One element of national security, the TSA, has a vast laboratory at Reagan National Airport devoted to studying systems and their usability, for protecting aviation travel. The three lead agencies charged with preventing attacks of *all kinds*, the CIA, FBI and NSA, in fact, the whole of the Intelligence Community, should have a similar space. With sufficient computer security it could be a *virtual space*. Agencies could commit their analysts, on rotation, to use the laboratory systems. Multiple analysts could work on the same problems, each using at least two systems, generating the information needed for the proposed analysis. This extends a pilot model called RDEC[6] that was developed a few years ago to place some experimental systems in or near the stream of real work being done by intelligence analysts.

We close with three specific proposals. (1) As a high national priority, establish a laboratory where the IC can study system performance, with real analysts and real problems. (2) In parallel, as quickly as possible establish research programs, perhaps using a cross-agency initiative similar to the ("Big Data is a Big Deal," 2012) focus. This permits multiple funding agencies to prioritize and support the most promising academic approaches to the problems of instrument design, and the development and validation of the underlying theories. Such pilot research is needed before justify testing those same instruments and theories in the crucible of the national effort to control, resist, and ultimately eliminate the threat of terrorism.

The laboratory could be "virtual" or "in the cloud." However, the second prong of the initiative may incline research universities more toward classified research, which could be advanced by a third initiative: (3) a program of term-long or semester-long work as "scholar in residence" at a suitable secure facility which is, itself, connected to the multi-agency virtual laboratory. Possible National Laboratory sites include: Brookhaven National Laboratory; Los Alamos; Lawrence Berkeley; Lawrence Livermore; Oak Ridge; and Pacific Northwest. Each serves a different part of the community, and would attract its own cohort of researchers.

In sum, there is great opportunity for the social sciences to advance national security, not only by theorizing, generalizing, and abstracting, but also by using established methods of applied social sciences to particularize, support, and accelerate development of the increasingly vital human⊗computer systems that protect the nation.

[5] The academic equivalent is to spend a lot of time "sharpening pencils."

[6] We did not find published discussions of RDEC. The principles discussed here have been used in a research analysis with real analysts and simulated problems in (Morse et al., 2004).

References.

Al-Maskari, A., & Sanderson, M. 2011. The effect of user characteristics on search effectiveness in information retrieval. Information Processing & Management 47(5), 719-729.

Big Data is a Big Deal [WWW Document], 2012. whitehouse.gov. URL https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal (accessed 6.7.17).

Cleverdon, C., 1967. The Cranfield tests on index language devices, in: Aslib Proceedings. MCB UP Ltd, pp. 173–194.

Hearst, M., Pedersen, J., Pirolli, P., Schutze, H., Grefenstette, G., Hull, D., 1995. Xerox site report: Four TREC-4 tracks, in: Proceedings of the Fourth Text Retrieval Conference.

Kantor, P., Boros, E., 2010. Deceptive Detection Methods for Effective Security with Inadequate Budgets: The Testing Power Index. Risk Anal. 30, 663–673. doi:10.1111/j.1539-6924.2010.01370.x

Kantor, P.B., 1988. National, Language Specific Evaluation Sites for Retrieval Systems and Interfaces. User-Oriented Content Based Text and Image Handling, in: RIAO Conference. MIT, pp. 138–147.

Morse, E.L., Scholtz, J, Kantor, P, Kelly, D, Sun, Y., 2004. ARDA Challenge Workshop 2004: An Investigation of Evaluation Metrics for Analytic Question Answering (No. NA). NIST.

Over, P., 2001. The TREC interactive track: an annotated bibliography. Inf. Process. Manag. 37, 369–381.

Saracevic, T., Kantor, P., 1988a. A study of information seeking and retrieving. II. Users, questions and effectiveness. J. Am. Soc. Inf. Sci. 39(3): 177-196.

Saracevic, T., Kantor, P., 1988b. A study of information seeking and retrieving. III. Searchers, searches, and overlap. J. Am. Soc. Inf. Sci. 39, 197–216.

Smith, C.L., Kantor, P.B., 2008. User adaptation: Good results from poor systems, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 147–154.

Sun, Y., 2005. Automatic assessment of non-topical properties of text by machine learning methods (Rutgers University Doctoral dissertation).

Sun, Y., Kantor, P.B., 2006. Cross-Evaluation: A new model for information system evaluation. J. Am. Soc. Inf. Sci. Technol. 57, 614–628.

Swets, J.A., W P Jr Tanner, Birdsall T G, 1961. Decision processes in perception. Psychol Rev. 68, 301–340.

Van Rijsbergen, C.J., 1974. Foundation of evaluation. J. Doc. 30, 365–373.